A Study of Tennis Score Evaluation Based on Logistic Regression and LSTM Neural Networks

Yuhao Cao *

School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China

* Corresponding Author Email: 3097226951@qq.com

Abstract. In order to study the performance of athletes in the competition, this paper establishes a comprehensive evaluation mechanism and prediction model through mathematical modeling. Firstly, this paper uses the clustering algorithm to categorize the data and establishes a score evaluation model according to the weights. Then, a momentum evaluation score calculation function was constructed and compared with the results of logistic regression algorithm. Finally, a multi-indicator data system was established to predict the results of the game using the LSTM neural network model. The experimental results show that the model proposed in this paper has a good prediction effect and provides suggestions to the participants.

Keywords: Clustering algorithm, logistic regression, LSTM neural network.

1. Introduction

In the game, it is necessary to develop a specific tactical system for players. Therefore, studying the evaluation mechanism and prediction model of athletes' performance in the game is of great significance and value for improving the quality of the game and cultivating excellent athletes [1]. Firstly, this paper selects 5 parameters representing athletes' stage performance status based on real game data, uses clustering algorithm to classify the data, and establishes a score evaluation formula based on the calculated weights of different parameters. Then, we provide a momentum evaluation score calculation function, based on the calculation function, simulate the total integral momentum of the match, combined with logistic regression algorithm, and compare the results. Finally, we established a data system under multiple indicators, analyzed the correlation of each parameter using the correlation function, and selected five parameters with evaluation characteristics. The LSTM neural network model was used to learn to predict the results of the game. The results show that the accuracy of the model is 0.95, and the prediction results are in good agreement with the actual results.

2. Score Evaluation Model

2.1. Selection of Indicators

To build a model for assessing the performance level of tennis players in a match, choosing suitable evaluation indicators is the basis for building the model. Therefore, we choose the following five indicators as the basic objects of the model. As shown in Table 1, "type" "+" and "-" in the third column indicate positive and negative indexes respectively.

Table 1. Evaluation indicators for the degree of performance in tennis matches

Indicators	Explanation	Type
ASTTM (y_1)	As the server of this tennis match	+
TTWP (y_2)	Total Tennis Winning Points	+
TNTWE (y_3)	Total number of tennis winning events	+
TCSVP (y_4)	Tennis court second victory probability	+
TNTTPL (y_5)	The number of times a tennis player loses	-

2.2. Weighting Model for Evaluation Indicators

In order to establish a mechanism for assessing the impact of first strike on winning and losing, these metrics need to be weighted. To avoid repetition and to make up for the lack of Entropy Weight method, we choose the method of the Power Average (PA) operator [2].

The Power Average (PA) Operator method uses power operations to provide an aggregation operator that allows parameter values to support each other in the aggregation process.

According to Table 1, we have this matrix:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{1}$$

Where each row of this matrix is all the indicators for a country.

Then perform the following three steps.

1. The formula for normalization is determined as:

$$x_{ij} = \frac{x_{ij}}{\max_j x_{ij}} \tag{2}$$

2. To quantify the support relationship between two indicators x_{ij} and x_{ik}, the relationship function is defined as follows.

$$R\left(x_{ij}, x_{ik}\right) = Ke^{-\alpha\left(x_{ij}-x_{ik}\right)^2} \tag{3}$$

Where $K \in [0,1]$ and $\alpha \geq 0$.

3. The weight vector is given by the following formula.

$$W = Pa(X_1, X_2, \cdots, X_m) = \frac{\sum_{i=1}^{m}\left((1+S(X_i))X_i\right)}{\sum_{i=1}^{m}\left(1+S(X_i)\right)} \tag{4}$$

Were,

$$S(X_i) = \sum_{h \neq i} R(X_i, X_h) \tag{5}$$

We can get a higher precision weight vector W through this model. Because PA operator can reflect the mutual supportiveness within the data. The results are shown in Fig 1.

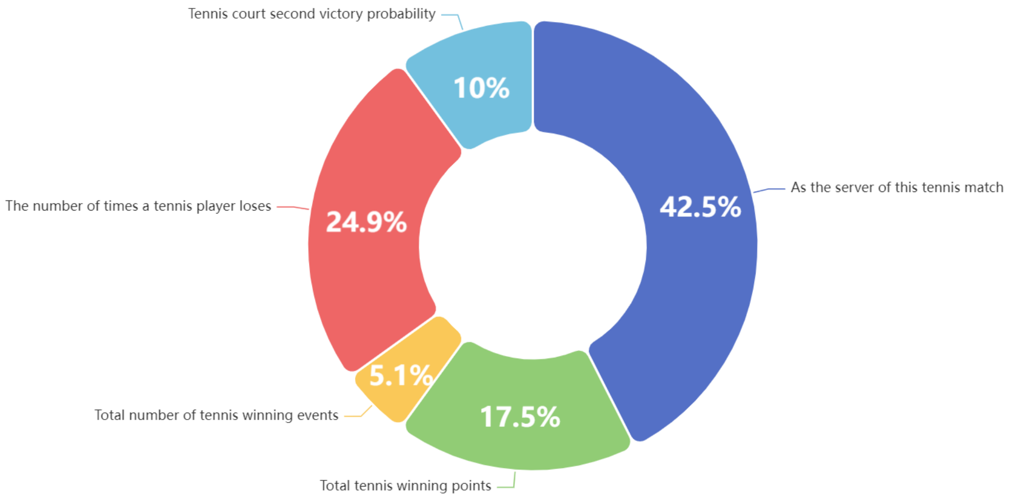

Figure 1. Weighting of indicators in the second system

2.3. Comprehensive Evaluation Model

From Fig 1, it is straightforward that the two evaluation functions are as follows.

For evaluation of the performance of tennis players, we have,

$$F = w_1 ASTTM + w_2 TTWP + w_3 TNTWE + w_4 TCSVP + w_5 TNTTPL \tag{6}$$

Overall evaluation of tennis players based on a closer assessment of their performance status, we have,

$$F' = \sum_{i=1}^{5} w_i' y_i \tag{7}$$

Where $W' = (w_1', w_2', w_3', w_4', w_5') = (0.325, 0.175, 0.051, 0.159, 0.276)$

2.4. Metric of Performance Degree for Players

The total score of each tennis player can be obtained from the above operations. Using the K-means clustering algorithm to cluster the scores of all tennis segments [3], the standard for the performance of tennis players is obtained. Meanwhile, a comprehensive standard was obtained using this method. The performance standards are shown in Fig 2.

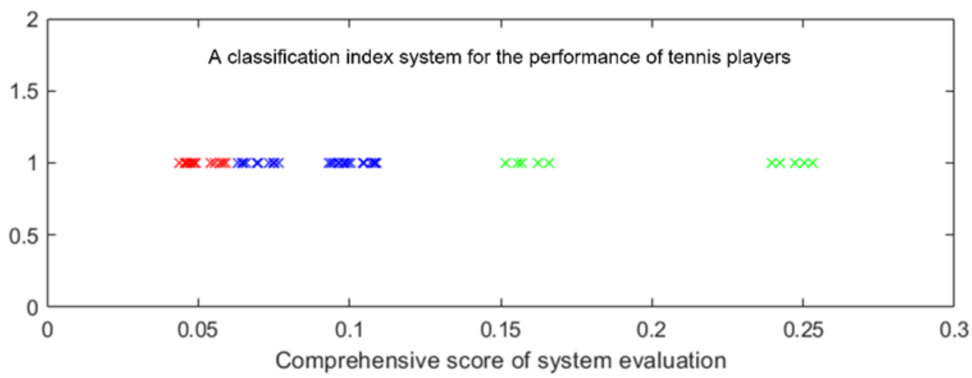

Figure 2. The results of K-means of two systems

The evaluation level of the performance of each tennis player is shown in Fig 3. We use the model in 4.3 to obtain any nation' s score. If the score is between 0.135 and 0.3, the performance is good.

If the score is between 0.07 and 0.135, the performance is not bad. If the score is between 0 and 0.07, the performance needs great improvement.

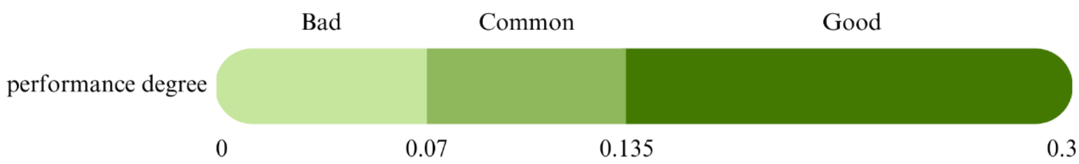

Figure 3. The threshold for evaluating the performance of tennis players

3. Momentum Assessment Model

3.1. Establish a Momentum Rating System

First, we establish a scoring system for momentum. Since momentum refers to the energy accumulated by athletes through winning or losing games during the competition, we can associate it with the probability of athletes winning over time in normal competitions, and first to quantify the momentum in the competition, we can build a model to evaluate the contribution of each scoring point to the player's momentum.

$$M_t = \alpha \sum_{i=1}^{t} P_i - \beta \sum_{i=1}^{t} L_i \quad (8)$$

We assigned two initial weights to the overall results as shown in Table 2 below.

Table 2. Weight assignment table

Weight indicators	Explanation	Initial value
α	The impact of scoring on momentum	0.63
β	The impact of losing points on momentum	0.37

At the same time, in order to characterize the actual competition trend, we calculate the change in momentum to make the results clearer, as shown below formula.

$$\Delta M_t = M_t - M_{t-1} \quad (9)$$

We obtained the following scoring probability calculation formula based on the two players in the entire field.

$$\delta_1 = \frac{M_{t-1}}{M_{t-1} + M_{t-2}} \quad (10)$$

Were δ_1 represents the scoring probability of the first tennis player, M_{t-1} represents the first tennis player's momentum quantification value at time point t, M_{t-2} represents the second tennis player's momentum quantification value at time point t, The scoring probability of another tennis player is calculated in the same way.

3.2. Logistic Regression Model with Parameter Correction

The following are the principles and steps of the logistic regression model [4].

1. Collect and prepare a dataset to ensure that the data includes input features and corresponding category labels.
2. Initialize model parameters, including weights and bias terms.
3. Calculate linear combinations: $z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$.
4. Convert linear combinations into probability values through logistic functions: $\sigma(z) = \frac{1}{1+e^z}$ ←.
5. Use a loss function to measure the difference between the model's output probability and the true label and perform calculations.

For better validation, the winning data is grouped according to player 1 and player 2 and labeled as 0,1 variables. If player 1 wins, player 1 is labeled as 1 and player 2 is labeled as 0. Then, based on two consecutive matches and three consecutive matches, the data is reorganized, and the original data is extracted to obtain a new analysis group, which is then regressed.

3.3. Momentum Assessment Scores and Regression Results

To eliminate the influence of randomness, we use a model to simulate the situation of 100 rounds in a single field and provide the results.

Under the momentum evaluation model, the momentum changes of two players in 100 matches are as follows Fig 4.

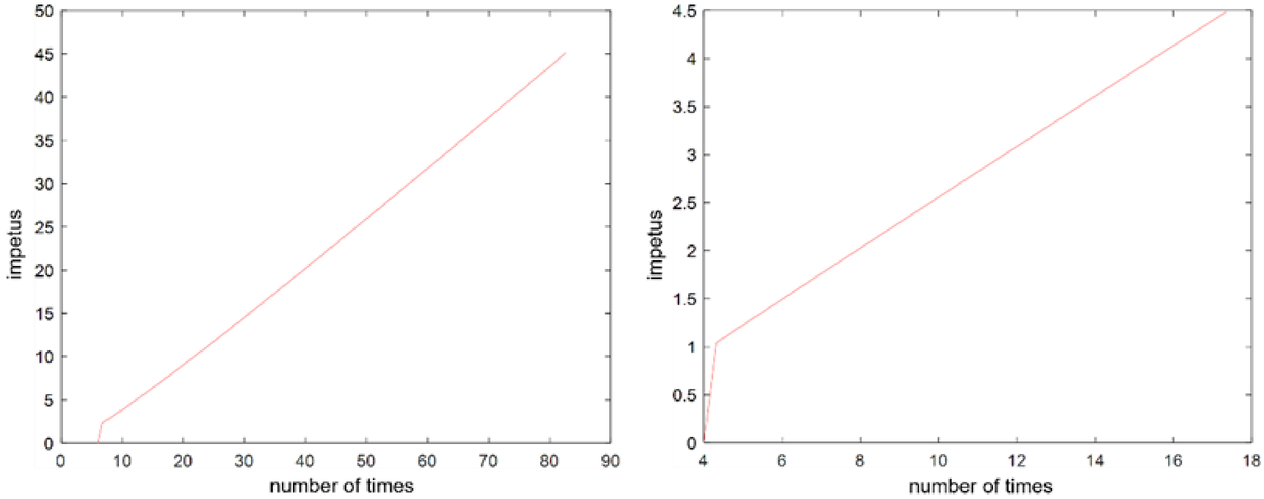

Figure 4. The momentum accumulation of two tennis players under momentum assessment

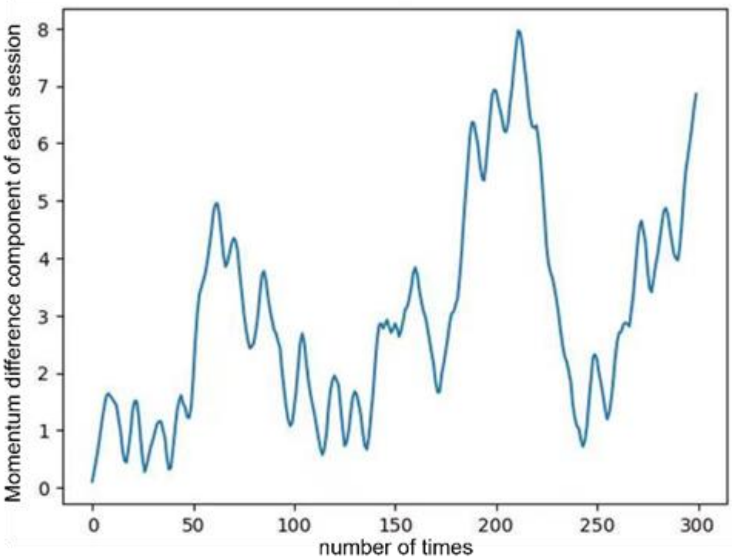

Figure 5. Momentum difference

As can be seen in Fig 4 and Fig 5, as one of the tennis players wins more matches during a match, they accumulate more momentum of their own. We only plotted the data for the number of wins and momentum, as shown in Fig 6. According to the attached data, the first player accumulated more momentum over the 100 matches, so the first player had a better chance of winning the last match, which is consistent with the actual results.

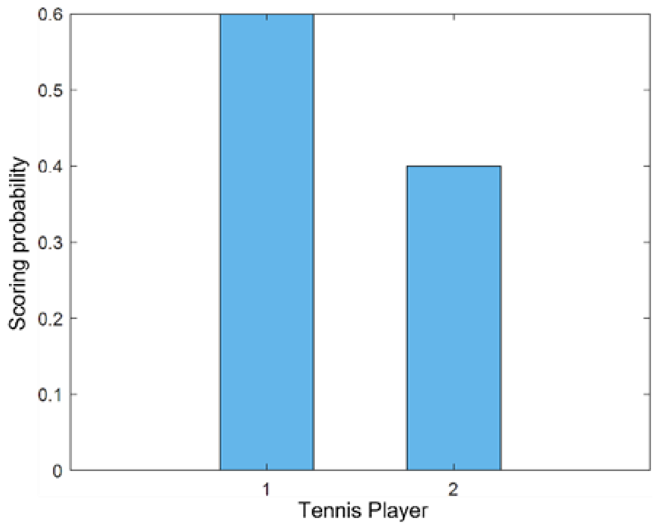

Figure 6. Revised overall scoring probability of two players

At the same time, we selected the winning matches from the actual data for regression analysis, as shown in the above figure. The overall error is small, and the results obtained are all the first tennis player to win, we also calculated the overall error of the regression, as shown in Fig 7. The overall error is small, and the fluctuation range is small.

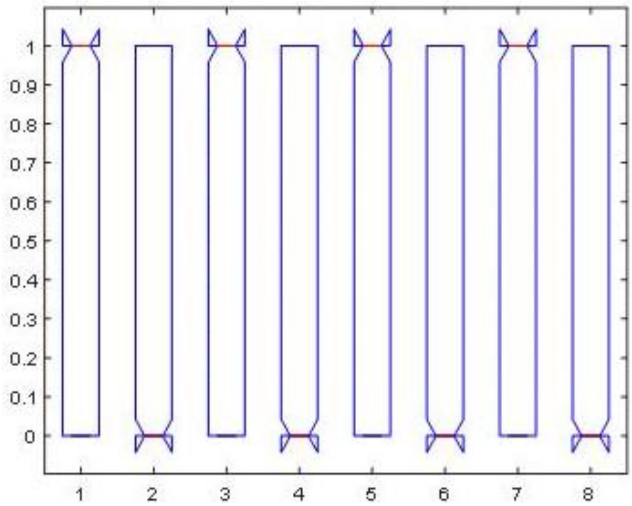

Figure 7. Logistic regression box diagram

4. Momentum Fluctuation Prediction Model

4.1. Adaptation of Memory Networks

The core technology of deep learning - LSTM is a new algorithm structure improved from RNN [5]. It not only has strong loop back characteristics but also can solve the problems encountered by traditional RNNs when facing complex tasks such as overfitting and underfitting and can to some extent overcome its shortcomings such as difficulty in capturing long-term correlations and dynamic changes [6].

The specific establishment mode is as follows.

1. Establish a forget gate: combining time t input x_t, control the memory state C_{t-1} of the previous time, and determine the old information to be forgotten and retained until the time t, as follows.

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{11}$$

2. Establish input gate: the sigmoid layer and tanh layer of the input gate jointly determine the new information added to the new cell state at time t. The calculation formula is as follows.

$$i_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_i\right) \tag{12}$$

4.2. Momentum Fluctuation Prediction and Model Validation

Based on the above analysis, we extracted and built a realistic network penetration model with 5 input layers and 1 output layer.

On this basis, we conducted a simulation test on the test set data using the network model, and the test results are shown in Fig 8. At the same time, the variance and error of the model calculation results were evaluated and simulated, and the validation results were obtained as shown in Fig 9, Fig 10 and Fig 11. The results show that the evaluation performance of the model is good, and the prediction variance of the actual data is low, so it can be considered to be a good prediction of the actual trend of changes in the use process.

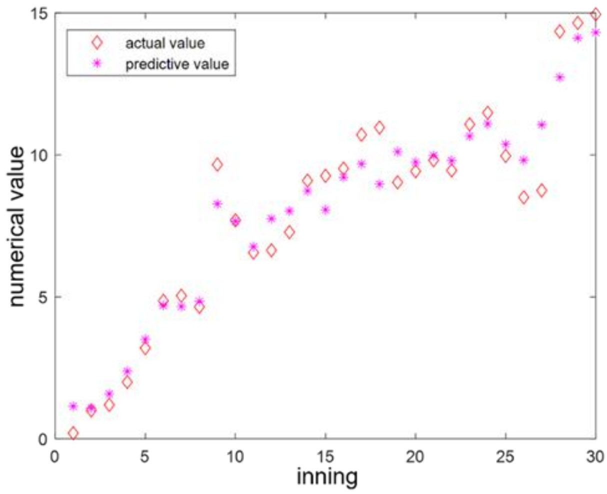

Figure 8. Comparison of test set network training prediction results

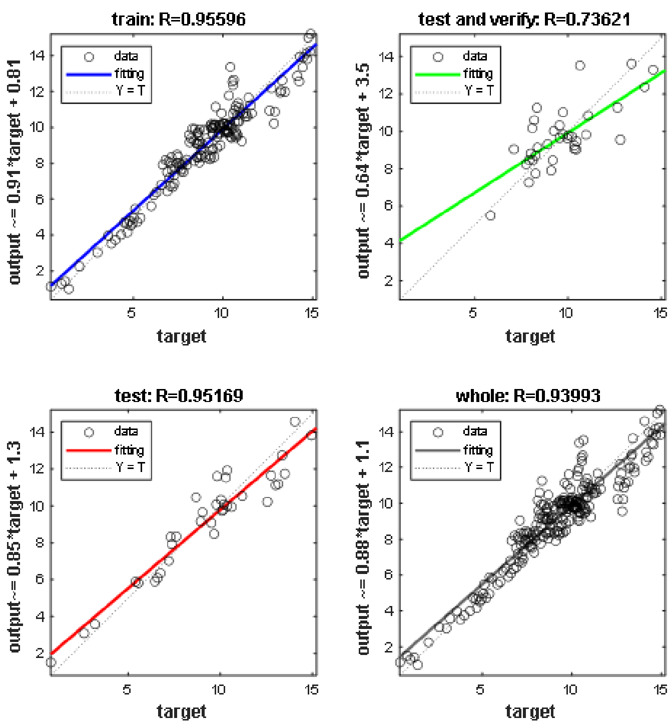

Figure 9. Network training regression plot

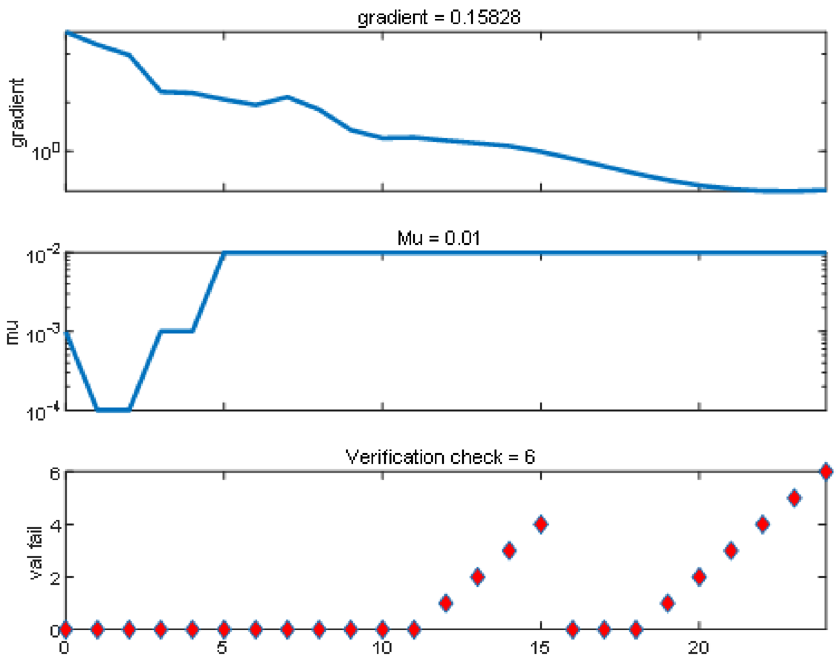

Figure 10. Network training state diagram

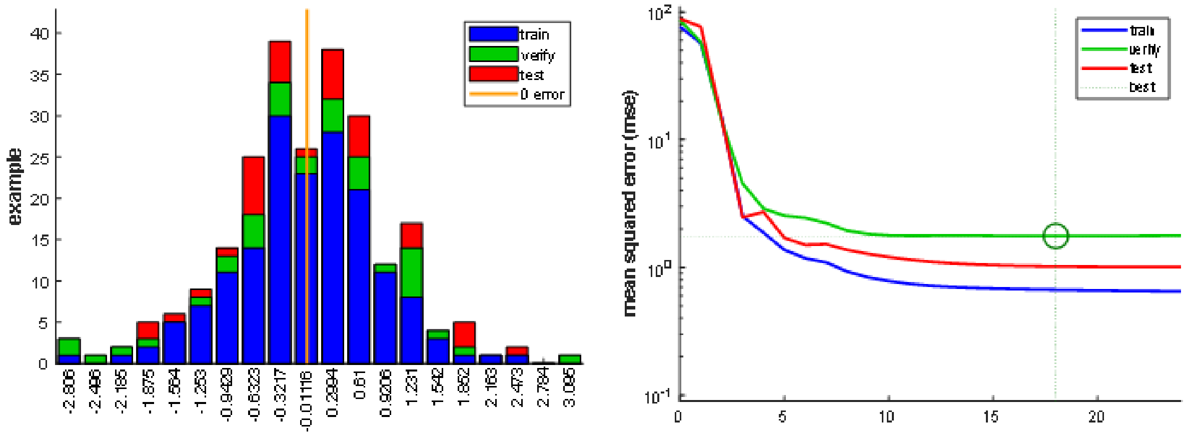

Figure 11. Error plot of network training results

5. Summary

In this paper, we studied the performance mechanism of athletes based on the existing data, evaluated the competitive trend of tennis players based on the performance mechanism, and established a big data network prediction model that can realize good data prediction. First, we used clustering algorithm to categorize the data and established a score evaluation formula based on the calculated weights of different parameters. Metrics for athletes to evaluate the range of scores were also provided in order to update the state data at different periods to obtain the evaluation values. Then, we provide a momentum evaluation score calculation function. Based on the calculation function, the matched total integral momentum was simulated and combined with a logistic regression algorithm to reproduce the results of the original dependent data, and the results were compared. We find that the probability of success of a player can be predicted using the momentum accumulation evaluation mechanism. Finally, an LSTM neural network model was learned using a test set to validate the model. The results show that the overall accuracy of the model is 0.95, and the prediction results are in good agreement with the actual results.

References

[1] Vitale, Jacopo A., et al. "Effect of a night game on actigraphy-based sleep quality and perceived recovery in top-level volleyball athletes." International Journal of Sports Physiology and Performance 14.2 (2019): 265 - 269.

[2] Chen, Tiedong, and Long Ye. "A novel decision-making method for selecting superintendent based on a q-rung dual hesitant fuzzy power partitioned Bonferroni mean operator." Symmetry 14.3 (2022): 590.

[3] Ghazal, Taher M. "Performances of k-means clustering algorithm with different distance metrics." Intelligent Automation & Soft Computing 30.2 (2021): 735 - 742.

[4] Park, Hyeoun. "An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain." Journal of Korean Academy of Nursing 43.2 (2013): 154 - 164.

[5] Rashid, Tarik A., Polla Fattah, and Delan K. Awla. "Using accuracy measure for improving the training of LSTM with metaheuristic algorithms." Procedia computer science 140 (2018): 324 - 333.

[6] Ijaz, Khalid, et al. "A novel temporal feature selection-based LSTM model for electrical short-term load forecasting." IEEE Access 10 (2022): 82596 - 82613.